

# **EXHIBIT 115**

AEO

**UNITED STATES DISTRICT COURT**  
**FOR THE NORTHERN DISTRICT OF CALIFORNIA**  
**SAN FRANCISCO DIVISION**

RICHARD KADREY, et al.,

*Individual and Representative Plaintiffs,*

v.

META PLATFORMS, INC.,

*Defendant.*

Case No. 3:23-cv-03417-VC

**REBUTTAL REPORT OF**  
**EMILY M. BENDER, PhD**  
**FEBRUARY 03, 2025**

input data, and suggests that this is important transformation. In fact, however, the value of the books to Meta is squarely in the selection and arrangement of the words contained therein. This is shown both by fundamental principles of linguistics and the way that Meta processes the data.

### **III. BACKGROUND**

#### **A. Methodologies from Corpus Linguistics**

17. The subfield of corpus linguistics approaches the study of the structure and use of language through the construction of collections of texts (corpora).<sup>8</sup> For a corpus to support scientific investigation, it must be representative of the language use under investigation. For example, if a research team is investigating formal versus informal language in some community, they might collect a sample of social media posts to represent informal language use and a sample of letters to the editor to represent more formal language use. In these paired corpora they could then look for differences in word choice and grammatical constructions to build up an understanding of what features of language count as “formal” and “informal” for that community. If the samples are too small, represent too few people or too few topics, then differences that are unrelated to the distinction in formality might overwhelm the kinds of differences the researchers are interested in.

18. The field of corpus linguistics has developed methodologies of corpus construction and documentation in order to support valid scientific investigations. These methodologies are necessary because of how linguistic artifacts — the texts that constitute corpora — are produced. When we speak or write, our word choices reflect the topic and content of what we are talking about, as well as the linguistic systems we participate in. Those linguistic systems include grammar (things like subject-verb agreement, how to construct conditionals, or how to express an indirect quote)<sup>9</sup> and associations between certain forms and certain social

---

<sup>8</sup> Kennedy, Graeme. (2014). *An introduction to corpus linguistics*. Routledge.

<sup>9</sup> Bybee, Joan L. (2006). From usage to grammar: The mind’s response to repetition. *Language*, 82(4), 711-733; Sag, Ivan A., Wasow, Thomas, & Bender, Emily M. (2003). *Syntactic theory: A formal introduction* (Vol. 92). Stanford, CA: Center for the Study of Language and Information.

identities or acts (like sounding “educated” or speaking “formally”).<sup>10</sup> Texts produced by people (as opposed to the synthetic output of LLMs) are thus rich in information that includes but goes beyond the information the author was directly talking about. They include information about the grammar of the language being used, and for anyone else who participates in the same linguistic community, information about the author’s identity and social presence.

19. If a corpus were chopped up into its individual letters and those letters were rearranged, it would cease to have value for most scientific investigations. A tiny amount of information might persist (these authors chose words with fewer ‘e’s in them than those authors), but the vast majority of it would go. The value of a text for communication between people and for scientific investigation is in the selection and arrangement of the words. Similarly, if the research goal is to build a system that can mimic text of a certain kind, such as an LLM, then the value of text for that project is in the selection and arrangement of the words.

#### **B. A Brief Overview of Language Models from a Linguistics Perspective**

20. Language models are an old type of technology, going back to the work of Claude Shannon in the 1940s.<sup>11</sup> Their fundamental task is to model the distribution of words in text: which words co-occur with which other words, and in which order. This is a useful component of technologies such as spell checkers, automatic transcription systems, and machine translation. Language models are useful in such technologies because they allow systems to rank different possible sequences of words are more or less likely.<sup>12</sup>

21. Today’s “large language models” (LLMs), including Meta’s Llama series of models, differ from those of the past in two respects. The first is scale: they are trained on much larger datasets and comprise much larger sets of weights so that they can take advantage of the

---

<sup>10</sup> Eckert, Penelope, & Rickford, John R. (Eds.). (2001). *Style and sociolinguistic variation*. Cambridge University Press.

<sup>11</sup> Schwartz, Oscar. (2019). Andrey Markov & Claude Shannon Counted Letters to Build the First Language-Generation Models. Internet: <https://spectrum.ieee.org/andrey-markov-and-claude-shannon-built-the-first-language-generation-models>, 5.

<sup>12</sup> Jurafsky, Daniel & Martin, James H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Online manuscript released January 12, 2025 <https://web.stanford.edu/~jurafsky/slp3/> This is the most up-to-date version of a widely used textbook in the field.

large datasets. The second is in how they used. While LLMs are still used in their original function of ranking sequences of words produced by some other process, they have gained prominence as synthetic text generation systems. In this mode, LLMs are used to repeatedly answer the question: “What is a word that is likely to come next, given the training data?”<sup>13</sup>

22. In any use case, language models are crucially dependent on their training data for their functionality. A language model used as part of an automatic transcription system will function very poorly in the transcription of newscasts if its training data does not include the names of people and places mentioned in the news. Similarly, an LLM being used to synthesize text will only be able to output text that looks like literary fiction if its training data includes sufficient examples of literary fiction.

23. A language model is a system for modeling the distribution of word forms (spellings of whole words or parts of words) in its “training” text. (The technical term for running the algorithm that produces the model over the input data is *training* and the data is called *training data*. This language is anthropomorphizing, but it is also the established term of art, so I will use it in this document.)

24. That is, a language model is a system for distinguishing more likely sequences of words or letters from less likely ones. The most basic language models are built by simply counting the frequency of individual words. This kind of model was behind the T9 system for writing text messages on phone keypads. If you type 4-6-6-3 that might have been 'home' or it might have been 'good'. The T9 system would rank these choices according to frequency of those words in its training data. The next level up looks at words in terms of the one to several preceding words. This kind of simple model is useful for spell checkers: a system might identify a word that is missing from its dictionary, determine what words in its dictionary are one or two letter exchanges away from what was typed, and then present those choices ranked according to

---

<sup>13</sup> Jurafsky, Daniel & Martin, James H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. Online manuscript released January 12, 2025 <https://web.stanford.edu/~jurafsky/slp3/>

how frequently they follow the immediately preceding word in the document (again in some training data).

25. Other use cases for language models include using them as components of systems for automatic transcription and machine translation. In these use cases, some other system component produces a set of candidate sentences (What might the audio signal be written as? What might that French string correspond to in English?) and the language model has the task of ranking those candidates according to what is most probable, given the information in its training data. For example, the English sentences *It is difficult to wreck a nice beach* and *It is difficult to recognize speech* sound very similar. If one or the other is spoken, both are likely to be presented as candidates to the language model in an automatic transcription system. Then, depending on the training data for that language model, one will be selected as the more probable. Another example is in machine translation. Turkish doesn't distinguish gender in the third person, so there is only one word which covers both *he* and *she*. Machine translation systems will frequently reveal the effect of their language models when translating from Turkish to English. The Turkish sentence for “he/she is a doctor” will get translated as *he is a doctor* and the Turkish sentence for “he/she is a nurse” will get translated as *she is a nurse*.<sup>14</sup> This is because the original sentence does not provide enough information for the translation model to definitively pick he or she, and so the bias encoded in the language model — the information about what words tend to go together in its training data — will kick in.

26. Earlier languages models were built directly out of counting sequences of words in a corpus: How frequent are doctor vs. nurse in the context *She is a \_\_\_\_*? Today's large language models function somewhat differently. The primary input is still large collections of text, but the language modeling software is designed to better capture the relationships between words and their contexts. In addition, between the initial trained model and the products that a person might interact with are other steps including some or all of:

---

<sup>14</sup> Caliskan, Aylin, Bryson, Joanna J., & Narayanan, Arvind. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

- a. Additional training based on ratings collected from people who evaluate system output. (“RLHF” or reinforcement learning from human feedback);
- b. Wrapper programs that add further information (“prompts”) to what the user might input;
- c. Pre- or post-processing guardrails that catch either inputs that the LLM provider would like to reject or outputs that need to be cleaned up;
- d. Systems for detecting inputs best handled by other software (such as calculators) and formatting the input for handing off to that software. These additions help build the illusion that large language models are “reasoning” engines, but the core of the LLM is still a model of the distribution of word forms in text.

**C. Language Models Model the Form of Language, and Only the Form**

27. It is very common to see claims of language models “understanding” language or otherwise having access to “meaning”. However, language models, like all machine learning models, are systems for modeling patterns in their training data. The training data for language models is sequences of *tokens* (the spellings of words or parts of words). If a person who can read the language in question were to look at this training data, it would appear that the language model were being given meaningful text (and thus information about the world or imaginary worlds). But the sequences of tokens are only meaningful to us as fluent readers of the language because we have the linguistic competence to understand how they are combined and what messages authors might be using them to convey.

28. To make this vivid, imagine a person with no knowledge of the Thai language all alone in the National Library of Thailand. Before this person arrived, all books with anything other than Thai writing were removed (no picture books, no mathematical equations, no translation dictionaries, etc). Even with unlimited time and access to everything ever written in Thai, there would be no way for that person to learn to read Thai — unless they could bring in external information, for example, by recognizing a particular Thai book as a translation of a

book they are already familiar with. Such outside information is not available to a language model.

29. The fundamental purpose of a language model is to represent the likelihood of given strings of tokens (words or word parts) given the data that was used to build it. Even when a language model is being used to synthesize output text, it is still fundamentally modeling such probabilities. This stands in stark contrast to how people use language — for myriad purposes most of which involve communicating with other people — even if the resulting collections of text might look similar. Whenever we are working with synthetic text from language models, it is important to remember that if it makes sense, it is because we are making sense of it.

**D. Relationship Between Data and Model in LLMs**

30. Training an LLM isn't just about turning words into numbers, though numerical representations are involved. In particular, the algorithms that produce the trained models are designed to represent words in terms of what other words they co-occur with — and those representations are vectors of numbers. These vectors are made up of numbers and they are not directly counts of what other words have been observed nearby, but rather the result of complex calculations pertaining to which other words the word has co-occurred with elsewhere in the text and which words are in the vicinity of the current instance of the word. Regardless of the complexity (and computational cost) of these calculations, the only source of information that is input into the system is the arrangement of words in the training data.

31. The whole value of the system rests on the numbers being a good representation of which words go where in the training data. If this weren't so, we could build equally good LLMs by taking words in the dictionary and repeating them random numbers of times to create the training text, or taking the training text, chopping it up into individual words, shuffling or alphabetizing the words, and then training the model on that. This methodology is not pursued, because it would not be effective. In other words, training data is valuable to builders of these



**A. Dr. Ungar Argues That Large Collections Of Written Text Are Necessary For LLM Development, But Particular Works Are Unimportant**

79. Throughout the report, Dr. Ungar argues that for the purpose of creating LLMs, large datasets of high quality text are necessary, in order to be able to output text on a wide variety of domains (what Dr. Ungar refers to as “capabilities”). Examples can be seen in paragraphs 63, 66, 90, 91, and 147, inter alia.

80. Dr. Ungar also argues that long-form text is useful for working with LLMs that can take in long passages of text (paragraph 239), that a sample of each work is not enough (paragraph 263), and that it is important to use “high-quality” data, such as can be found in genuine books (paragraph 216). In sum, it is clear from this report that training data is immensely important to the development of the Llama models. I find that this position is inconsistent with his argument that particular parts of that training data, especially parts that are long, high quality texts, are unimportant.

**B. The Preprocessing Done on Input Data Does Not Remove Information About The Selection and Arrangement of Words Therein**

81. Dr. Ungar argues that the preprocessing and processing done on the input data turns it into something substantially different.

82. Regarding preprocessing, he writes, “This tokenization process modifies the format and structure of the original input data, making it understandable for neural networks and large language models, but removing all human-readable meaning of the underlying words” (paragraph 130). But the tokenization process (mapping from word parts to numerical representations of those word parts) still preserves the identity and order of the word parts. For a person to be able to read the sequence once it had been translated this way, they would need to have the key, that is the mapping between word parts and numbers. At this stage in the process as described by Dr. Ungar in paragraphs 69-72 and 130, the mapping is fully reversible, and it is irrelevant that a person without the key would find the sequence meaningless. The authors’ selection and arrangement of words is still very much preserved.

83. Dr. Ungar goes on to describe, at a high level, how the input texts are used in producing model weights through the process of training. At this point, the process becomes non-reversible, as he states (paragraph 132). However, this lack of reversibility does not negate the importance of original texts to the process of LLM development. The fact remains that the value that Meta found in the books used in the training data lies in the selection and arrangement of the words in the text. I elaborate on this point below.

84. Dr. Ungar also argues that the processing done on the input data constitutes a “profound transformation” (paragraph 137). I note that Dr. Ungar does not provide a definition of “transformation” (or “transformativeness”), nor is he a legal scholar qualified to opine on the question of how the law sees the relationship between a set of training data and the weights derived from it.

85. In another discussion of “transformativeness” (also paragraph 137), Dr. Ungar asserts that LLM outputs aren’t based on the inputs. But if that were true, then constructing an LLM wouldn’t require an enormous training data set. In short, there are three components of an LLM: its model architecture, its training regime, and its dataset. Meta announces that the Llama models are “open-source”, but one of these three components is not released nor thoroughly described, and that’s the training data. Furthermore, the models are valued for their outputs (the text they can be used to synthesize). In my opinion, this shows that the outputs are very much dependent on the inputs.

**C. The Value of Books to Meta is Squarely in the Selection and Arrangement of the Works Contained Therein**

86. There is also plenty of evidence in Dr. Ungar’s report that the value of the books as training data is precisely the selection and arrangement of the words contained therein. Dr. Ungar writes that LLM’s “capabilities instead derive from the structure of the network and the connections/weights, which through training, enable the network to recognize patterns and relationships in the data” (paragraph 30). Without the words as arranged by the authors of the texts, there would be no patterns or relationships for the model to represent.

87. Dr. Ungar also describes a process of deduplication, which is important to the effective training of LLMs (paragraph 209). The importance of deduplication implies the value of many, varied, quality texts. If the choice and arrangement of words didn't matter, deduplication wouldn't matter, either, since repeated sequences would be as valuable as varied sequences.

88. Dr. Ungar emphasizes the enormous amount of computing power being used to process the training data (paragraph 140). This is a lot of processing, to be sure, and it means that the resulting artifact is not just a database of the input data. However, it also shows that Meta is finding a lot of value in the input data. The processing would be meaningless without it.

89. Finally, Dr. Ungar uses “noise texts” as a point of comparison in his experiments and motivates them as follows: “To provide additional comparison, a second set of models are further pretrained on “noise” texts—repetitive and nonsensical texts that do not help the model learn any meaningful information (e.g., repetitions of the word “the”)” (paragraph 152 point E). Presupposed in this characterization is the idea that the high quality texts included in the training data are high quality because they contain not nonsensical strings, but words carefully and intentionally arranged by authors.

90. As discussion in Section II, Part 1, the methodology of corpus linguistics underscores that the import of a text lies not just in its length, but in the word choice and grammatical structures it includes. These are the patterns that the LLMs are designed to represent and serve as the basis for their functionality in outputting text that mimics all of the genres in their training data. Without data that displays the patterns, no amount of training would result in an LLM that can represent the patterns.

## **IX. CONCLUSION**

91. In conclusion, I find that the opinions that Dr. Ungar presents as unsupported for the following reasons:

AEO


92. His report is threaded through with misleadingly anthropomorphizing language about LLMs. This presentation is misleading as to the functionality of LLMs, setting them up as something other than systems for representing the distribution of word forms in text.

93. On the basis of this anthropomorphization, he develops a misleading presentation of the notion of “generalization”, wherein the LLMs are agents doing something with their input data and this somehow means that the input texts were not material to the construction of the models.

94. Dr. Ungar’s choice of the MMLU benchmark rests on unsound scientific practice, as that benchmark does not measure what it purports to measure: language models do not understand anything, so a benchmark designed for language models to measure their “language understanding” is meaningless. Dr. Ungar furthermore doesn’t establish any other foundation for the benchmark in order to make it meaningful in his experiments.

95. Dr. Ungar’s anthropomorphizing language displaces accountability. In using anthropomorphizing language, which places the Llama models as the agent of actions, Dr. Ungar displaces accountability away from Meta and to the models. In any deliberations about the development and use of “AI” systems, it is important to maintain clarity on who is taking action.

96. Finally, Dr. Ungar’s treatment of the value of written works to the development of large language models is internally inconsistent and unsupported. He argues both that “high quality” written works like books are necessary for LLM development, and (simultaneously) that any given written work is not important. These two positions are inconsistent. Even the largest dataset of written works is made up of individual works and cannot be created without them. Dr. Ungar also describes the processing of the input data as if it constituted “transformation”. This does not change the fact that the value to Meta of the works was their original selection and arrangement of words.



---

Emily M. Bender, PhD